*Mol. Biol. Med.* (1988) **5**, 107–122

# Cloning and Characterization of a cDNA Encoding Human Galactose-1-phosphate Uridyl Transferase

**Juergen K. V. Reichardt and Paul Berg**

*Department of Biochemistry*
*Stanford University School of Medicine*
*Stanford, CA 94305, U.S.A.*

*(Received 25 January 1988)*

**Summary** We report the cloning and characterization of a cDNA that encodes a functional human galactose-1-phosphate uridyl transferase (GALT). The cDNA is 1400 bases in length and encodes a 43,000 $M_r$ protein. The cloning strategy involved the identification of short peptide sequences conserved between the homologous enzymes from *Escherichia coli* and yeast, and the construction of oligonucleotide pools corresponding to the conserved patches. These patches of conserved amino acids tend to be conserved in humans as well.

## Introduction

Mammalian cells metabolize galactose (Gal) after its conversion to glucose-1-phosphate (G-1-P)† *via* three well-characterized reactions (Segal, 1983). First, galactokinase (GALK) catalyzes the phosphorylation of Gal by ATP to produce galactose-1-phosphate (Gal-1-P). Second, galactose-1-phosphate uridyl transferase (GALT) transfers the uridyl moiety of uridine diphosphoglucose (UDPG) to Gal-1-P to form uridine diphosphogalactose (UDP-Gal) and G-1-P. UDP-Gal-4-epimerase (GALE) regenerates UDPG by inverting the orientation of the hydroxyl group at position C-4 in the galactosyl portion of UDP-Gal.

Galactosemia results from a defect in any one of the three enzymes needed to metabolize galactose. The most prevalent form of galactosemia, however, is attributable to a deficiency of GALT (McKusick no. 23040: McKusick, 1986). This form of the disease is inherited as an autosomal recessive trait and homozygotes occur with a frequency of 1 in 50,000 newborns without any particular ethnic or geographic variation (Levy & Hammersen, 1978). Very little or no GALT activity is detectable in the tissues of such affected individuals. The GALT gene, which has been mapped to human chromosome 9 band p13 (Meera Khan & Robson, 1978) has not been cloned

†Abbreviations used: G-1-P, glucose-1-phosphate; GALT, galactose-1-phosphate uridyl transferase; Gal-1-P, galactose-1-phosphate; UDPG, uridine diphospho-glucose; UDP-Gal, uridine diphospho-galactose; GALE, UDP-Gal-4-epimerase; SSCE, 150 mM-NaCl, 15 mM-sodium citrate, 1·5 mM-EDTA; kb, $10^3$ base (pair); bp, base pair; hprt, hypoxanthine phosphoribosyltransferase.

0735–1313/88/020107+16 $03.00/0

and, therefore, nothing is known about the molecular nature of the genetic defect in GALT galactosemics.

The corresponding GALT genes in *Escherichia coli* (Lemaire & Mueller-Hill, 1986) and *Saccharomyces cerevisiae* (Tajima *et al.*, 1985) have been cloned and their nucleotide sequences are known. The translated amino acid sequences of the bacterial and yeast proteins show only a 40% similarity; however, there are several short contiguous stretches of five or six amino acids each that occur in similar positions of the two proteins. Lacking any information about the amino acid sequence of the human GALT protein, we reasoned that the conserved sequences in the *E. coli* and *S. cerevisiae* enzymes might also occur in the human GALT protein. Accordingly, we used degenerate oligonucleotide pools corresponding to the conserved amino acid stretches as hybridization probes to isolate a cDNA clone that encodes the entire human GALT protein. Here we report the isolation and characterization of the cloned cDNA sequence and its expression after transfection into cultured primate and human cells.

## Materials and methods

### (1) *Materials*

Oligonucleotides were synthesized by an Applied Biosystems 380B DNA synthesizer, purified by polyacrylamide gel electrophoresis and labeled with [$\gamma$-$^{32}$P]ATP (NEN; 6000 Ci/mmol) and phage T4 polynucleotide kinase.

The plasmid pNN75, which contains the yeast *GAL7* gene (St John & Davis, 1981) was obtained from Ronald Davis (Stanford University). The hamster–human hybrid cell lines and the CHO-K1 parent were from Carol Jones (Denver, CO; Jones *et al.*, 1980). The simian virus 40 (SV40) transformed human galactosemic fibroblasts (GM639) were obtained from the Institute of Medical Research, Camden, N.J., SV40 transformed normal fibroblasts (VA13) were obtained from Gilbert Chu (Stanford University) and IM9 lymphoid cells were from Brian Pontius (Stanford University).

### (2) *Methods*

Standard recombinant DNA techniques were performed as described (Maniatis *et al.*, 1982). A size-selected pcD-cDNA expression library in *E. coli* HB101 (Okayama & Berg, 1983) was screened at a density of 30,000 colonies per 150 mm Petri dish with the mixed oligonucleotide probes ($10^6$ cts/min per ml) in 6 × SSCE, 0·1% (w/v) SDS, 3 × Denhardt's medium, 50 mM-sodium phosphate (pH 7·0) at 37°C. Blot transfers of RNA denatured with formaldehyde (Northerns) and DNA (Southerns) were made to GeneScreen Plus (NEN), according to the manufacturer's recommendations, and hybridized with $10^6$ cts/min per ml of hexamer-labeled cDNA probe (Feinberg & Vogelstein, 1983). For nucleotide sequence analysis, the cDNA was cloned into BlueScript (Stratagene) and nested deletions were created with exonuclease III (Henikoff, 1984); individual deletions were rescued with R408 helper phage and sequenced on both strands by the dideoxy chain termination method (Sanger *et al.*, 1977). The sequence data were managed with the Bionet system and amino acid sequence comparisons used release 12·0 of the protein database.

Electroporations were performed as described (Chu *et al.*, 1987) using 20 μg of plasmid DNA and 200 μg of tRNA as carrier. RNA was prepared from cells lysed in guanidinium thiocyanate/Sarkosyl by centrifugation through a CsCl step gradient (Chirgwin *et al.*, 1979). Genomic DNA was isolated after lysis of cells in 150 mM-NaCl, 10 mM-Tris (pH 8), 10 mM-EDTA, 0·2% (w/v) SDS and digestion with 0·1 mg proteinase K/ml for 5 to 16 h at 50°C. DNA was collected by precipitation with ethanol after equilibrium centrifugation in 0·95 g CsCl/ml, 1 mg ethidium bromide/ml (typically in a Beckman 80 Ti rotor at 400,000 *g* for 15 h). DNA from human sperm was released with 100 mM-2-mercaptoethanol prior to digestion with proteinase K.

IM9 human lymphoblastoid cells were grown in RPMI 1640 medium (Gibco) supplemented with 10% (v/v) horse serum (HyClone), penicillin and streptomycin. The hamster–human cell hybrids and the hamster parent were grown in Ham's F12 (Gibco) medium supplemented with 10% horse serum, penicillin and streptomycin and the GM639, VA13 and COS7 cells were grown in DME (Gibco) containing the same supplements.

Cell extracts used to assay GALT were prepared by first washing the cells 3 times in ice-cold Tris-buffered saline (pH 7·4), then the cells were lysed in 10 mM-Tris · HCl (pH 8·5), 10 mM-dithiothreitol by freeze–thawing once followed by sonication for 2 min and centrifugation at 15,000 *g* for 10 min in the cold. GALT activity was measured by the disappearance of UDPG according to published procedures (Mellman & Tedesco, 1965). Corrections were made for any reaction in the absence of added Gal-1-P or in the absence of cell extracts. Protein was determined with the BioRad dye-binding assay according to the manufacturer's instructions with bovine serum albumin as internal standard.

## Results

Our aim was to isolate the human gene encoding the GALT protein. However, we possessed neither nucleotide sequence data about the gene and its mRNA, nor any amino acid sequence information about the protein, nor a means for screening or selecting cells that express the GALT gene. This led us to explore a strategy that relies on the evolutionary conservation of amino acid sequences in proteins with identical functions in such distantly related organisms as *E. coli* and *S. cerevisiae*. A computer program generated alignment of the amino acid sequence of the galactose-1-phosphate uridyl transferase from *E. coli* (galT) (Lemaire & Mueller-Hill, 1986) with the sequence of the corresponding protein in *S. cerevisiae* (GAL7) (Tajima *et al.*, 1985) revealed an overall sequence similarity of about 40% (Fig. 1). There were numerous stretches in which five or more amino acid sequences were identical between the two proteins.

### (1) *Isolation of GALT cDNA*

Assuming that one or more of these sequences also occurred in the human GALT protein, we synthesized totally degenerate oligonucleotide pools from six of these

***E.coli* and yeast GALT**

```
E.coli      MT    QFNPVDHPHRRYNPLTGQWILVSPHR LSPWQAQETPAKQVLPAH
Yeast       MTAAEEFDFSSHSHRRYNPLTDSWILVSPHRAKRPWLGQEAAYKPTAPLY
Consensus   MT----F----H-HRRYNPLT--WILVSPHR---PW--QE---K---P--
oligo

            DPDCFLCAGNVRVTGDKNPDYTGTYVFTNDFAALMSD      TPDAPESH
            DPKCYLCPGNKRATGNLPNRYESTYIFPNDYAAVSDQPILPQNDSNEDNLN
            DP-C-LC-GN-R-TG-----Y--TY-F-ND-AA------------------

            DPLMRCQSARGTSRVICFSPDHSKTLPELSVAALTEIVKTWQ       EQ
            NRLLKVQSVRGNCFVICFSPNHNLTIPQMKQSDLVHIVNSWQALTDDLSRE
            --L---QS-R----VICFSP-H--T-P------L--IV--WQ---------
                          AAAAAA

            TAELGKTYPWVQ FENKGAAMGCSNPHPHGQIWANSFLPNEAEREDRLQKE
            ARENHKPFKYVQIFENKGTAMGCSNLHPHGQAWCLESIPSEVSQELKSFDK
            --E--K----VQ-FENKG-AMGCSN-HPHGQ-W-----P-E---E------
                               BBBBBB

            YFAEQKSPMLVDYVQRELADGSRTVVETEHWLAVVPYWAAWPFETLLLPKA
            YKREHNTDLFADYVKLESREKSRVVVENESFIVVVPYWAIWPFETLVISKK
            Y--E-------DYV--E----SR-VVE-E----VVPYWA-WPFETL-----
                                                    CCCCCC

            HVLRITDLTDAQRSDLALALKKLTSRYDNLFQCSFPYSMGWHGAP  FNGE
            KLASISQFNQMAKEDLASILKQLTIKYDNLFETSFPYSMGIHQAPLNATGD
            ----I---------DLA--LK-LT--YDNLF--SFPYSMG---AP----G-
                                      DDDDD

            ENQHWQLHAHFYPPLLRSATVRKFMVGYEMLAETQRDLLTAEQAAERLR A
            ELSNSWFHMHFYPPLLRSATVRKFLVGFELLGEPQRDL ISEQAAEKLRNL
            E------H-HFYPPLLRSATVRKF-VG-E-L-Q-QRDL---EQAAE-LR--
                     FFFFFFFFFFFFFFF                 EEEEE

            VSDIHFRESGV
            DGQIHYLRQRL
            ---IH------
```

**Figure 1.** A comparison of the amino acid sequences in galactose-1-phosphate uridyl transferases from *E. coli* and *S. cerevisiae*. The amino acid sequences, shown in the 1-letter code, are aligned using the Bionet GENALIGN program. The consensus sequence is the sequence in common between *E. coli* and yeast. The oligonucleotide pools described in Table 1 were based on the regions labeled A through F.

stretches for use as hybridization probes with human cDNA libraries (Table 1 and Fig. 1). Probes A, B, D and E were labeled with $^{32}P$ at their 5′ ends and used to screen 90,000 colonies comprising a pcD-cDNA library with a nominal insert size of 1·5 to 2·0 kb (Okayama & Berg, 1983). This size-cut was chosen because the human GALT protein (from both red blood cells and placenta) has been reported to be about 45,000 $M_r$ (Williams *et al.*, 1982) and, therefore, would require a coding sequence of at least 1·2 kb in the mRNA. Expecting that the existence of 5′ and 3′ untranslated regions would make the mRNA somewhat larger, we chose the sub-library with the 1·5 to 2·0 kb inserts.

**TABLE 1**

E. coli-*yeast conserved probes*

| Name | Amino acid sequence | Degeneracy of nucleotide sequences | $T_d^{min}$ (°C) |
|---|---|---|---|
| A | VICFSP | 768 | 42 |
| B | AMGCSN | 512 | 44 |
| C | WPFETL | 128 | 46 |
| D | YDNFL | 512 | 40 |
| E | EQAAE | 128 | 44 |
| F | HFYPPLLRSATVRKF | $10^8$ | 72 |

The minimum dissociation temperatures ($T_d^{min}$) were estimated for 1 M-Na$^+$ concentration assuming the lowest possible G+C contents (Suggs *et al.*, 1981).

The initial hybridizations identified 21 clones that appeared to hybridize to varying degrees with the four oligonucleotide pools. However, on rescreening each of the 21 putative positive clones by Southern blotting with probes A, B, D and E, only one, pcD-GALT, was consistently positive when hybridized with the four probes. Indeed, this clone, also hybridized with the oligonucleotide pools A through E at 44°C and with pool F at 44°C in buffer containing 20% formamide.

After digestion with *Xho*I endonuclease (Fig. 2(a), lane 1) or *Bam*HI endonuclease (Fig. 2(a), lane 2), the putative pcD-GALT plasmid yields three discernible fragments: the largest corresponds to the expected size of the cloning plasmid when cleavages are made at the restriction sites flanking the cDNA insert (Fig. 3); the sum of the lengths of the two *Xho*I fragments and of the two *Bam*HI fragments is about 1·6 kb. This indicates that the cDNA segment has at least one *Xho*I and *Bam*HI restriction site.

Each of the oligonucleotide pools (A to F) hybridized principally to the larger of the two insert fragments (Fig. 2(b), lanes 1 and 2). None of the probes hybridized to the lane containing *Bam*HI endonuclease-digested pcD-hprt (Jolly *et al.*, 1983) (Fig. 2(b), lane 3). As anticipated, each of the probes hybridized to the DNA fragment encoding the yeast GAL7 protein (St John & Davis, 1981) (Fig. 2(b), lane 4). These data indicate that the cloned cDNA in pcD-GALT is related to the corresponding coding sequences of *galT* in *E. coli* and *GAL7* in yeast.

Using the cDNA insert as a hybridization probe to screen a $\lambda$ phage cDNA library (Okayama & Berg, 1985) we found two positive clones in $10^6$ recombinants. Unless

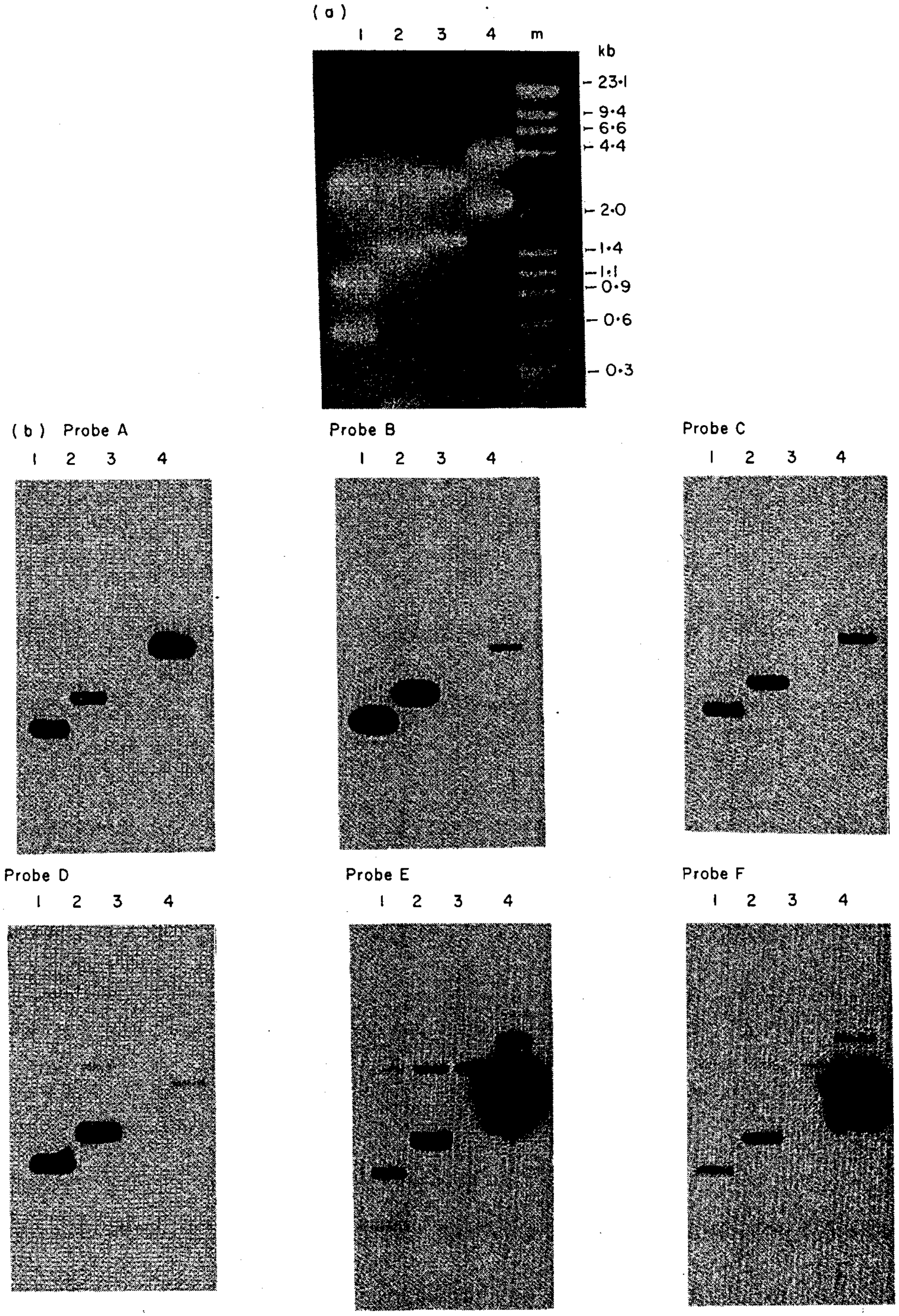

**Figure 2.** Hybridization of oligonucleotides to pcD-GALT. (a) The ethidium bromide-stained gel; (b) autoradiograms of the various restriction endonuclease digestions after hybridization with the labeled oligonucleotide pools shown in Table 1. Lanes 1, pcD-GALT cleaved with *Xho*I endonuclease; lanes 2, pcD-GALT cleaved with *Bam*HI; lanes 3, pcD-hprt (Jolly *et al.*, 1983) cleaved with *Bam*HI endonuclease; lanes 4, pNN75 (contains the yeast *GAL7* gene; St John & Davis, 1981) cleaved with *Sal*I endonuclease.

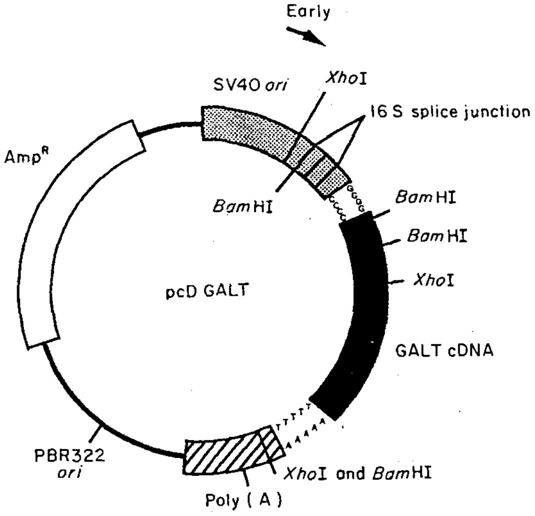

**Figure 3.** The structure of pcD-GALT. The positions of relevant restriction sites are shown. The filled region corresponds to the GALT cDNA; the stippled region contains the SV40 early region promotor, *ori* and an intron derived from the SV40 16 S late mRNA (Okayama & Berg, 1983); the hatched segment contains the SV40 early region polyadenylation signal.

there is a bias against the cloning of this sequence in the phage vector, this implies that the mRNA encoding GALT is a minor species.

### (2) *Characterization of GALT cDNA*

After summing up the sizes of the cDNA insert fragments and correcting for the lengths of vector sequence that occur in both the *Xho*I and *Bam*HI restriction fragments, we estimate that the cDNA insert in pcD-GALT is about 1·4 kb in length. The smaller fragments produced in each digestion are not detected in the hybridizations because all the oligonucleotide probes correspond to nucleotide sequences downstream from the *Xho*I and *Bam*HI restriction sites (Fig. 3, and the cDNA sequence in Fig. 5). Since the cloned cDNA hybridizes to a 1·4 kb mRNA (Fig. 4), the cDNA probably contains the entire coding sequence.

Nucleotide sequencing of the cDNA insert revealed that it has a single long open reading frame which can encode a 43,000 $M_r$ protein (Fig. 5); this agrees well with the 45,000 $M_r$ size estimated for the human red cell and placenta enzymes (Williams *et al.*, 1982). The open reading frame begins with an ATG codon 72 bp from the 5′ end of the cDNA and continues for 380 codons to a TGA codon located 97 bp from the 3′ end. A consensus polyadenylation signal, AATAAA (Proudfoot & Brownee, 1976) begins 79 bp beyond the TGA codon and 18 bp upstream from the poly(A) tail. There is an out-of-frame ATG triplet 62 bp 5′ to the beginning of the long open reading frame, but this is followed almost immediately by an in-frame termination codon and a second in-frame terminator closer to the protein's initiator codon. Neither the first nor the presumptive initiator ATG fits the proposed consensus sequences that surround most start codons for eukaryotic proteins (Kozak, 1984).

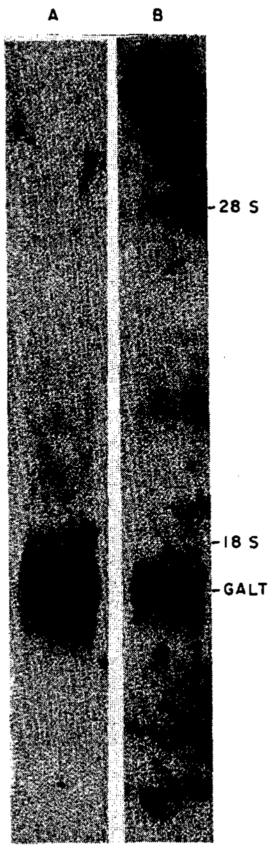

**Figure 4.** Northern blots of GALT mRNA: 5 $\mu$G of poly(A)$^+$ RNA from IM9 lymphoid cells (lane A) or VA13 fibroblasts (lane B) were fractionated on a 1·2% formaldehyde agarose gel and hybridized as described in Materials and methods.

### (3) *Transfection of mammalian cells with pcD-GALT results in the formation of GALT activity*

The pcD vector system was developed to allow for the expression of the cloned cDNAs after transfection into appropriate mammalian cells (Okayama & Berg, 1983). Transcription is initiated from the SV40 early region promoter, splicing occurs at an intron placed between the promoter and the coding sequence, and polyadenylation occurs either within the cDNA segment or at a polyadenylation signal 3′ to the cDNA segment (Fig. 3). The presence of the SV40 origin of DNA replication allows for amplification of the DNA in cells that express a functional large T antigen (Gluzman, 1981).

To determine if transfection of mammalian cells with pcD-GALT would lead to the formation of GALT, we carried out the following experiment. Either the T-antigen-producing COS7 cell line or an SV40-transformed human fibroblast cell line derived from a galactosemic individual (GM639) were transfected with one of the following

**TABLE 2**
*Expression of GALT after transfections with pcD-GALT*

| Cell line | Plasmid | Specific activity (units/μg)† |
|---|---|---|
| COS | pSV2-neo | 1·0 |
| COS | pcD-hprt | 0·7 |
| COS | pcD-GALT | 40·0 |
| GM 639 | pSV2-neo | <0·05 |
| GM 639 | pcD-hprt | 0·05 |
| GM 639 | pcD-GALT | 1·7 |

† 1 unit of GALT activity catalyzes the Gal-1-P dependent disappearance of 1 μmol of UDPG/min (Mellman & Tedesco, 1965).

three plasmid DNAs: the putative pcD-GALT, pcD-hprt (Jolly *et al.*, 1983) and pSV2-neo (Southern & Berg, 1982). Extracts prepared two days following electroporation were assayed for GALT activity as described in Materials and methods (Table 2). Quite clearly, GALT activity is induced only in cells transfected with the pcD-GALT plasmid, the increase in activity in COS cells being 40 to 60-fold over that in the cells transfected with the two control plasmids. Similar results were obtained with the galactosemic cell line, the increase there being 30-fold. The 20-fold difference in the level of GALT activity achieved in the transfections with the COS *versus* GM639 cell lines is probably due to the amplification of the transfected plasmid DNA in COS cells (Tooze, 1980).

### (4) *The GALT gene*

The GALT gene has been mapped to human chromosome 9 band p13 (Meera Khan & Robson, 1978). We detect a single 12 kb band in *Eco*RI endonuclease digests of human DNA using the GALT cDNA as a hybridization probe (Fig. 6, lane D). A weakly cross-hybridizing 9 kb band can be seen in a similar digest of the DNA from the CHO-K1 cell line (Fig. 6, lane C). Hamster–human hybrids carrying human chromosome 9 have both the hamster and human specific bands (Fig. 6, lanes A and B). Hybrid J640-51 contains human chromosomes 9 and 22 while J640-63 harbors chromosomes 9, 13, 18 and 21 (Jones *et al.*, 1980). The reduced intensity of the human band in the two hybrid cell lines (Fig. 6, lanes A and B) is probably the result of chromosome loss. These findings indicate that sequences corresponding to the GALT cDNA reside on chromosome 9. Experiments being carried out in collaboration with Uta Francke's laboratory are intended to localize the chromosomal position more precisely.

The 12 kb genomic GALT segment has been cloned and parts of its sequence have been determined. However, it contains only the 0·4 kb at the 3′ terminus of the cDNA. Thus, the gene probably extends over considerably more than 12 kb, but fragments that contain sequences encoded in the 5′-proximal 1 kb of the cDNA have not been identified.

## GALT cDNA

```
gtattatccaATGCAGtgagatcgcctgtggattggcttcgaaatatcgaaataggatccc

ccggtggcctcATG TCG CGC AGT GGA ACC GAT CCT CAG CAA CGC CAG
           met ser arg ser gly thr asp pro gln gln arg gln

CAG GTG CAG AGG CGG ACC CGA GCA GCA ACC TTC CGG GCA AAC GAC
gln val gln arg arg thr arg ala ala thr phe arg ala asn asp

CAT CAG CAT ATC CGC TAC AAC CGG CTG CAG GAT GAG TGG GTG CTG
his gln his ile arg tyr asn arg leu gln asp glu trp val leu

GTG TCA GCT CAC CGC ATG AAG CGC CCT GCA GGG TCA AGT GGA GCC
val ser ala his arg met lys arg pro ala gly ser ser gly ala

CCA GCT TCT GAA GAC AGT GCC CCG CAT GAC CCT CTC AAC CTC TGT
pro ala ser glu asp ser ala pro his asp pro leu asn leu cys

GTC ACC TCT GTG TCC TGG GGA TCC GAG CCA AGG AGA GGT GAA TCC
val thr ser val ser trp gly ser glu pro arg arg gly glu ser

CAG TAC GAT AGC ACC TTC CTG TTT GAC AAC GAG CTT CCA GCT CTG
gln tyr asp ser thr phe leu phe asp asn glu leu pro ala leu

CAG CTG ATG CCC CAG TCC AGG ACC AGT GAT CAT CCC TTT TCA AGC
gln leu met pro gln ser arg thr ser asp his pro phe ser ser

AAG TCT GCT CGA GGA GTC TGT CAG GTC ATG TGC TTC CAC CCT GGT
lys ser ala arg gly val cys gln val met cys phe his pro gly

CGG ATG TCA CGC TGC CAC TCA TGT CGG TCC CTG AGA TCC GGG CTG
arg met ser arg cys his ser cys arg ser leu arg ser gly leu

TTG TTG ATC GAA TGG GCC TCA GTC ACA GAG GAG CTG GGT GCC CAG
leu leu ile glu trp ala ser val thr glu glu leu gly ala gln

TAC CCT TGG GTC GAG ATC TTT GAA AAC AAA GGT GCC ATG ATG GGC
tyr pro trp val glu ile phe glu asn lys gly ala met met gly

TGT TCT AAC CCC CAC CCC CAC TGC CAG GTA TCG GGC CAG AGT TTC
cys ser asn pro his pro his cys gln val ser gly gln ser phe

CTG CCA GAT ATT GCC CAG CGT GAG GAG CGA TCT CAG CAG GCC TAT
leu pro asp ile ala gln arg glu glu arg ser gln gln ala tyr

AAG AGT CAG CAT GGA GAG CCC CTG CTA ATG GAG TAC AGC CGC CAG
lys ser gln his gly glu pro leu leu met glu tyr ser arg gln

AGC TAC TCA GGA AGG AAG TCT GGT CCT AAC AGT GAG CAC TGG TTA
ser tyr ser gly arg lys ser gly pro asn ser glu his trp leu
```

```
GTA CTG GTC CCC TTC TGG GCA ACA TGG CCC TAC CAG ACA CTG CTC
val leu val pro phe trp ala thr trp pro tyr gln thr leu leu

GTT CCC GTC GGC CAT GTG CGG CGG CTA CCT GAG CTG ACC CCT GCT
val pro val gly his val arg arg leu pro glu leu thr pro ala

GAG CGT GAT GAT CTA GCC TCC ATC ATG AAG AAG CTC TTG ACC AAG
glu arg asp asp leu ala ser ile met lys lys leu leu thr lys

TAT GAC AAC CTC TTT GAG ACG TCC TTT CCT ACT TCA TGG GCT GGC
tyr asp asn leu phe glu thr ser phe pro thr ser trp ala gly

ATG GGG CTC CCA CAG GAT CAG AGG CTG GGG CAA TTG GAA CCA TTG
mrt gly leu pro gln asp gln arg leu gly gln leu glu pro leu

GCA GCT GCA CGT CAT TAC TAC CGT GCG CTG CGC TCT GCC ATG TCG
ala ala ala arg his tyr tyr arg ala leu arg ser ala met ser

GAA ATT AAT GGT TGG CTA CGA AAT GCT TGC GAG GCT CAG AGG GAC
glu ile asn gly trp leu arg asn ala cys glu ala gln arg asp

TAC CCT GAG CAG GCT GCA GAG AGA CTA AGG GCC ATT CCT GAG GTT
tyr pro glu gln ala ala glu arg leu arg ala ile pro glu val

CCA TTA CAC CTC GGG CCA GAA GAC CAG GGA GAC ACC AAC CAT CGC
pro leu his leu gly pro glu asp gln gly asp thr asn his arg

CTG ACC ACG CCG ACC ACA GGG CCT TGA atcctttttcttttcaacagtct
leu thr thr pro thr thr gly pro

tgctgaattaagcagaaagggcttcgaatcctcgctgaattggagatatagcattaataa

aactgtgcatcacaa......
```

**Figure 5.** Nucleotide sequence of the GALT cDNA. The long open reading frame begins with the ATG (shown in outline) codon following nucleotide 72 and ends with a TGA codon (shown in outline) near the 3′ end. Also shown are the positions of the *Bam*HI and *Xho*I restriction sites indicated in Fig. 3. The short upstream reading frame is italicized and the polyadenylation signal is shown in boldface. Both open reading frames are in capital letters.

## Discussion

In preparing hybridization probes to identify human cDNA clones encoding galactose-1-phosphate uridyl transferase (GALT), we assumed that even in organisms separated by long evolutionary distances, proteins with identical functions would share amino acid homologies. A comparison of the homologous enzymes from *E. coli* and *S. cerevisiae* revealed 12 patches of five or more identical amino acids at corresponding locations. Six of the common sequences were used to synthesize pools of all possible oligonucleotide sequences; the choice of the six was almost arbitrary although the more carboxy-terminal sequences were preferred in order to maximize the chances of finding clones having at least the 3′ end of the mRNA.

The cloned GALT cDNA is very nearly the same size as the mRNA which it detects in Northern blots. Moreover, its sequence probably encodes the entire GALT amino

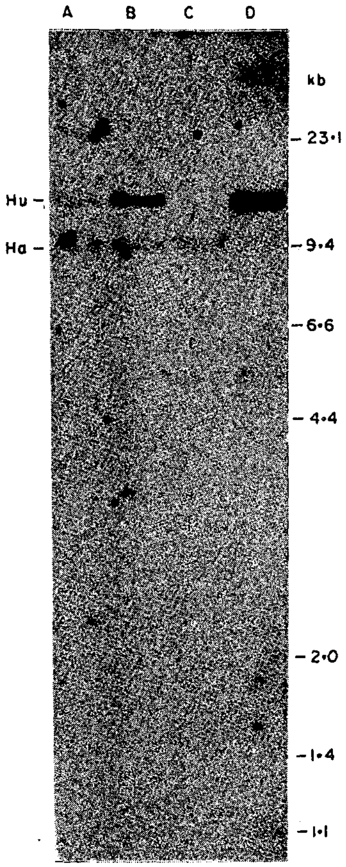

**Figure 6.** The GALT cDNA hybridizes to sequences in human chromosome 9. 10 $\mu$g of genomic DNA were digested with *Eco*RI endonuclease and the digests were electrophoresed through a 0·8% agarose gel, transferred and hybridized as described in Materials and methods. Lane A, CHO–human hybrid J640-63 (carrying human chromosomes 9, 13, 18 and 21); lane B, hybrid J640-51 (with human chromosomes 9 and 22); lane C, parent cell line CHO-K1; and lane D, human germline DNA. The human (Hu) and CHO (Ha) GALT bands are indicated.

acid sequence since there is a substantial increase in GALT activity following transfection of pcD-GALT into COS cells or into cells lacking GALT. The sequene of the cDNA itself is undistinguished. The coding sequence begins with an ATG triplet located 72 nucleotides beyond the 5′ end of the cDNA and continues for 1140 nucleotides to a TGA termination signal 97 nucleotides upstream from the 3′ end. The first ATG in the cDNA is ten nucleotides from the 5′ end but its reading frame is interrupted by two in-frame termination codons 5′ to the beginning of the GALT coding sequence.

**Figure 7.** Amino acid sequence comparison of GALT from human, *E. coli* and *S. cerevisiae*. The format for the 3 sequences and their alignment are as described for Fig. 1. The consensus sequence designates those residues that are conserved in the 3 enzyme species. Note that only regions D and E are conserved in all 3 organisms.

# *E.coli* yeast and human GALT

```
Human         MSRSGTDPQQRQQVQRRTRAATFRANDHQHIRYNRLQDEWVLVSAHR MKR
E.coli                         MT  QFNPVDHPHRRYNPLTGQWILVSPHR LSP
Yeast                          MTAEEFDFSSHSHRRYNPLTDSWILVSPHRAKRP
Consensus     ----------------------F----H-H-RYN-L---W-LVS-HR----
oligo

              PAGSSGAPASEDSARHDPLNLCVTSVSWGSEPRRGESQYDSTFLFD
              W QAQETPAKQVLPAHDPDCFLCAGNVRVTGDKNPDYTGTYVFTNDFAALM
              WLGQQEAAYKPTAPLYDPKCYLCPGNKRATGNLNPRYESTYIFPNDYAA V
              ----------------DP------------------------F--------

                NELPALQLMPQSRTSDHPFSSKSARGVCQVMCFHPGRMSRCHSCRSLRS
              SD       TPDAPESHDPLMRCQSARGTSRVICFSPDHSKTLPCLSVAAL
              SDQPILPQNDSNEDNLNNRLLKVQSVRGNCFVICFSPNHNLTIPQMKQSDL
              --------------------------RG---V-CF-P--------------
                                             AAAAAA

              GLLLIEW       ASVTEELGAQYPWVEIFENKGAMMGCSNPHPHCQVSG
              TEIVKTWG       EQTAELGKTYPWVQ FENKGAAMGCSNPHPHGQIWA
              VHIVNSWQALTDDLSREARENHKPFKYVQIFENKGTAMGCSNLHPHGQAWC
              ------W------------E-------V--FENKG--MGCSN-HPH-Q---
                                                  BBBBBB

              QSFLPDIAQRE ERSQQAYKSQHGERLLMEYSRQSYSGRKSGPNSEHWLVL
              NSFLPNEAEREDRLQKEYFAEQKSPMLVDYVQRELADGSRTVVETEHWLAV
              LESIPSEVSQELKSFDKYKREHNTDLFADYVKLESREKSRVVVENESFIVV
              ----P-----E----------------------------------E-----

              VPFWATWPYQTLLVPVGHVRRLPELTPACRDDLASIMKKLLTKYDNLFETS
              VPYWAAWPFETLLLPKAHVLRITDLTDAQRSDLALALKKLTSRYDNLFQCS
              VPYWAIWPFETLVISKKKLASISQFNQMAKEDLASILKQLTIKYDNLFETS
              VP-WA-WP--TL-------------------DLA---K-L---YDNLF--S
                   CCCCCC                                DDDDD

              FP TSWAGMGLPQDQRLGQLEPLAAARHYYRALRSA MSEINGWLRNACEA
              FPYSMGWHGAP  FNGEENQHWQLHAHFYPPLLRSATVRKFMVGYEMLAET
              FPYSMGIHQAPLNATGDELSNSWFHMHFYPPLLRSATVRKFLVGFELLGEP
              FP--------------------------Y---LRSA---------------
                                        FFFFFFFFFFFFFFFF

              QRD  YPEQAAERLR AIPEVPLHLGPEDQGDTNHRLTTPTTGP
              QRDLLTAEQAAERLR AVSDIHFRESGV
              QRDL ISEQAAEKLRNLDGQIHYLQRL
              QRD----EQAAE-LR-----------------------------
                     EEEEE
```

Fig. 7.

Southern blots of human genomic DNA digested with several different restriction enzymes and hybridized with the cloned GALT cDNA reveal only a single fragment. A 12 kb *Eco*RI fragment has been cloned but its sequence contains only the cDNAs 3′-terminal 0·4 kb separated by three introns. Thus far, no alterations have been detected in the genomic DNA of ten GALT galactosemic individuals when their Southern blots have been annealed with the GALT cDNA.

A comparison of the amino acid sequences of the *E. coli, S. cerevisiae* and human GALT enzymes reveals that there is only 25% overall similarity (Fig. 7). Nevertheless, there are four patches of at least five identical amino acids that occur at corresponding places in the three proteins (Fig. 7). Two of these patches occur in region B, one at position D and the other at E. The human sequence in region A differs by two amino acids from that of the *E. coli* and yeast proteins; one of the differences in region A is methionine in place of isoleucine, possibly due to a single base change. Conservative changes within the adjacent region C also account for the differences between the human, *E. coli* and yeast sequences; in this region, the human GALT sequence contains a phenylalanine residue instead of tyrosine and, in the adjoining patch, the human sequence has adjacent tyrosine and glutamine residues instead of the phenylalanine and glutamic acid sequence that occurs in *E. coli* and yeast. Curiously, in region F, where the *E. coli* and yeast GALT sequences share a stretch of 15 identical amino acids, there are only five matching amino acids in the human protein. And, where *E. coli* and yeast sequences share two patches of eight amino acids at their 5′ ends, there is only a spotty similarity in the human sequence.

The approach described here should prove useful for isolating protein-coding sequences for which there is neither nucleic acid nor protein sequence data available to guide the construction of hybridization probes. Indeed, the cDNA encoding the mammalian M2 subunit of ribonucleotide reductase was identified by hybridization with mixed oligonucleotide probes based upon two patches of shared amino acid sequence between the Herpes simplex 1, Epstein-Barr virus and clam enzymes (Thelander & Berg, 1986). Subsequently, the yeast and mammalian M2 subunits were found to be 60% similar in their amino acid sequences (Elledge & Davis, 1987). In a forthcoming paper we shall describe several additional examples in which sequence conservation between the same proteins from evolutionarily distant organisms provides a useful approach to the isolation of their related genes.

## Acknowledgments

This work was supported by Public Health Service grant GM-13235 from the National Institute of General Medical Sciences.

## References

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18**, 5294–5299.

Chu, G., Hayakawa, H. & Berg, P. (1987). Electroporation for the efficient transfection of mammalian cells with DNA. *Nucl. Acids Res.* **15,** 1311–1326.

Elledge, S. J. & Davis, R. W. (1987). Identification and isolation of the gene encoding the small subunit of ribonucleotide reductase from *Saccharomyces cerevisae*: DNA damage-inducible gene required for mitotic viability. *Mol. Cell. Biol.* **7,** 2783–2793.

Feinberg, A. P. & Vogelstein, B. (1983). A technique for radiolabelling restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **137,** 266–267.

Gluzman, Y. (1981). SV40-transformed simian cells support the replication of early SV40 mutants. *Cell* **23,** 175–182.

Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28,** 351–359.

Jolley, D. J., Okayama, H., Berg, P., Esty, D. C., Filpula, D., Bohlen, P., Johnson, G. G., Shively, J. E., Hunkapiller, T. & Friedman, T. (1983). Isolation and characterization of a full-length expressible cDNA for human hypoxanthine phosphoribosyltransferase. *Proc. Nat. Acad. Sci., U.S.A.* **80,** 477–481.

Jones, C., Kao, F. T. & Taylor, R. T. (1980). Chromosomal assignment for folylpolyglutamate synthetase to human chromosome 9. *Cytogenet. Cell Genet.* **28,** 181–194.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eucaryotic mRNAs. *Nucl. Acids Res.* **12,** 857–872.

Lemaire, H. G. & Mueller-Hill, B. (1986). Nucleotide sequence for the gal E and gal T gene of *E. coli. Nucl. Acids Res.* **14,** 7705–7711.

Levy, H. L. & Hammersen, G. (1978). Newborn screening for galactosemia and other galactose metabolism defects. *J. Pediat.* **92,** 871–877.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McKusick, V. A. (1986). *Mendelian Inheritance in Man,* 7th edit., The Johns Hopkins University Press, Baltimore, MD.

Meera Khan P. & Robson, E. B. (1978). Report of the committee on the genetic constitution of chromosome 9, Cytogen. *Cell Genet.* **22,** 106–110.

Mellman, W. J. & Tedesco, T. A. (1965). An improved assay of erythrocyte and leukocyte galactose-1-phosphate uridyl transferase: stabilization of the enzyme by a thiol protective reagent. *J. Lab. Clin. Med.* **66,** 980–986.

Okayama, H. & Berg, P. (1983). A cDNA cloning vector that permits the expression of cDNA inserts in mammalian cells. *Mol. Cell. Biol.* **3,** 280–289.

Okayama, H. & Berg, P. (1985). Bacteriophage lambda vector for transducing a cDNA library into mammalian cells. *Mol. Cell. Biol.* **5,** 1136–1142.

Proudfoot, N. J. & Brownee, G. G. (1976). 3′ Non-coding region in eukaryotic messenger RNA. *Nature (London)* **263,** 211–214.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci., U.S.A.* **74,** 5463–5467.

Segal, S. (1983). Disorders of galactose metabolism. In *The Metabolic Basis of Inherited Disease* (J. B. Stanbury *et al.*, eds.), pp. 167–191, McGraw-Hill Book Company, New York, NY.

Southern, P. J. & Berg, P. (1982). Transformation of mammalian cells to antibiotic resistance with a bacterial gene under control of the SV40 early region promoter. *J. Mol. Appl. Genet.* **1,** 327–341.

St John, T. P. & Davis, R. W. (1981). The organization and transcription of the galactose gene cluster of *Saccharomyces. J. Mol. Biol.* **152,** 285–315.

Suggs, S. V., Hirose, T., Miyake, T., Kawashima, E. H., Johnson, M. J., Itakura, K. & Wallace, R. B. (1981). Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences. *ICN-UCLA Symp. Mol. Cell. Biol.* **23,** 683–693.

Tajima, M., Nogi, Y. & Fukasawa, T. (1985). Primary structure of the *Saccharomyces cerevisiae GAL7* gene. *Yeast* **1,** 67–77.

Thelander, L. & Berg, P. (1986). Isolation and characterization of expressible cDNA clones encoding the M1 and M2 subunits of mouse ribonucleotide reductase. *Mol. Cell. Biol.* **6,** 3433–3442.

Tooze, J. (1980). Editor of *DNA Tumor Viruses*, 2nd edit., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Williams, V. P., Helmer, G. R. & Fried, C. (1982). Human galactose-1-phosphate uridyl transferase: purification and comparison of the red blood cell and placental enzyme. *Arch. Biochem. Biophys.* **216,** 503–511.